To treat or not to treat: the historical source before the input

Härtel, Reinhard

Historical Social Research, Vol. 14 —1989 — No. 1, 25-38

To Treat or not to Treat: The Historical Source Before the Input

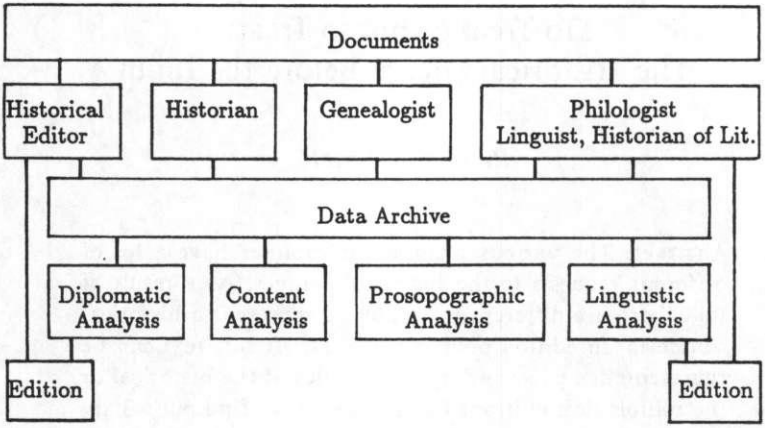
*Reinhard Härtel**

Abstract: The various historical disciplines have a lot of different requests to the historical sources. As a result of that there are different ways how to represent a historical source in an edition or in a data-base. A full-text can be represented, e.g., according to the rules of the historical or the philological edition. The problem is to find out a standard of full-text input which is acceptable for the whole scientific community. It is more likely that standardization of formalized documentation (concerning the details of text-treatment) can ensure the exchangeability and general usability of machine-readable sources than standardization of text-treatment can do. A formalized documentation could also be used for computer-supported producing of critical editions according to different rules on the basis of one and the same literal transcription.

1. Introduction

Historical sources are known to be questioned by various scientists in very different ways depending on the specific problems (1). Archivists, e.g., take a strong interest above all in doing inventary or repertory work; editors rather concentrate on problems of textual tradition and criticism, on the possibilities of typographical arrangement and on working out indices. The historians in the narrow sense of the word are first and foremost preoccupied with analysing the contents of texts and those doing diplomatics or similar studies will probably be keen on the inner structure of texts. And moreover if the historical sciences mentioned are to pay tribute to paleographers or linguists there will further be enormous demands involved. Graph 1 illustrates a few possibilities if the numerous possible combinations between those who enter data and those who use them.

* Address all communications to: Reinhard Härtel, Universität Graz, Forschungsinstitut für Historische Grundwissenschaften, Körblergasse 20, A - 8010 Graz, Austria.



It is evident that certain rules have been developed in several historical disciplines according to which a full text is to be transcribed and edited. There is a wide range of transcriptions from exact copying of letters and interpunction to a considerable normalization. As a rule one and the same text does not serve only the editor and his closest professional colleagues, but it also serves scholars of the most differing historical disciplines. Therefore it seems to be very important on the one hand to find a standard which does not demand too much extra-work from anybody, because too many colleagues would have to do work which is useless for them. On the other hand this standard cannot be so simple that a lot of information which is important for the majority of users gets lost. In the beginning, of course, we have not only to ask »To treat or not to treat«, but also »How to treat«.

Some colleagues seem very cautious in judging the chances of standardizing the representation as well as the necessary documentation of texts. In fact the progress in standardizing has been unsatisfactory at least till now. This is clearly shown by the proceedings of the standardization-conferences in Graz (2) and Paris (3). And there are doubts even in a more narrow range of application, e.g. dictionary, machine-readable summary, or analysis of structural units in the text. Such doubts might have been inforced through the fact that the types of written sources greatly vary between chronicles and inscriptions, and from country to country and from period to period. In addition to that we shall probably have to work for quite a long time with texts which are nothing else but conventional editions which have only been made machine-readable afterwards. And those editions have been arranged according to wide-ranging principles, which we

do not even know sometimes. As we can see, the problem of text-treatment does not only rise from problem-orientated input, but also from a mere full-text input.

An excessively great importance lies with the presentation of special projects scientists are occupied with. And, according to the respective participants and programs different projects and different problems are discussed, and that is why each time the results are incomplete and each time incomplete in a different way. As a result of that the problems of the paleographers have not been considered yet in the general discussion about standardization of text-input in a suitable way. Frequently the discussion on standardization is confined to study description. But also in the case of source-input description technical problems are predominant (coding systems etc.). The full-text treatment *before* the input is far too often neglected (4).

First of all a short analysis of the various requests which the historical disciplines can have in respect to historical texts is presented (5). So after seeing the different requests we have to consider a way of text-treatment which guarantees at least a certain level of usability of exchanged full-texts by the colleagues of other disciplines. The result will be that a noticeable progress in exchangeability and usability is rather possible by standardization of documentation, and not so much by standardization of text-treatment itself. Certainly there is the problem which has been discussed from time to time that some scientists would not like the meticulous publication of their principles, which would offer more possibilities for some professional »faultfinders«. In the end it will be pointed out that there is a glimmer of hope in a way, that the standardization of documentation could encourage standardization of text-input itself. So the headline-question »To treat or not to treat« will be answered, respectively also the question »How to treat«.

In order to avoid misunderstandings it should be stressed: we discuss fundamental problems of standardizing which lie before the use of the computer and therefore beyond all technical details. But of course, all these problems are connected with the computer, and so we always have the computer in view.

2. The Various Ways of Text-Treatment

Now first of all to the analysis of the requests which the historical disciplines can have, and to the respective solutions concerning the representation of texts, but not by giving a long list of special disciplines, because such a list would constantly change and will therefore be incomplete soon. It cannot be avoided to repeat some well-known facts in the beginning.

In any case the demands on a text can only concern four criteria, and it is evident that besides these four criteria there is nothing else to study in a written text: firstly its material contents, secondly its structure (both logical and linguistic), thirdly the graphic realisation (both the visual arrangement and the singular graphemes; concerning this point we are especially interested in handwritings), and finally the phonemes behind the graphemes. Moreover the various purposes of the singular scientists are irrelevant. Editors, historians (in the widest sense of the word) and lexicographers cannot study other things in a given text than its contents etc., as has been mentioned above. For our purpose we can recommend a little change: we will deal with the phonemes before we come to the graphemes. A short representation of the different types of treatment, which correspond to the four criteria mentioned, will not be completely useless, because it will be the basis of some suggestions concerning standardization.

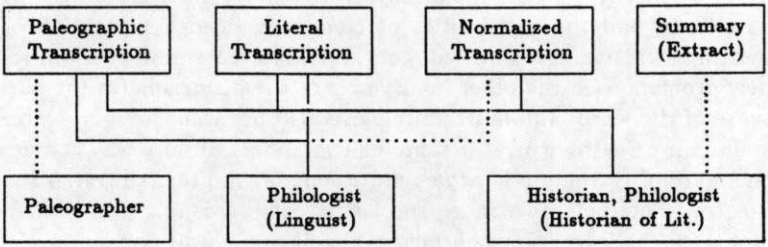
As to the contents it is obvious that a well-done summary in whatever language often can suffice as a basis for whatever research. But the machine-readable summary offers one of the most difficult problems for standardization, and it is also obvious that this is not part of our topic. It will be difficult, however, to avoid any loss of information when texting such a summary, but nevertheless a summary will often suffice for most purposes of archivists, librarians and historians (in a wide sense of the word). There will remain, however, the problems of special terms. Such problems can only be solved by a transcription.

As to the structure we have to distinguish between logical structure (occasionally expressed also by graphic arrangements or shown by the change of writers) on the one hand and linguistic structure on the other hand. As to the logical structure, an income list, e.g., can be set up according to people or possessors. As to the linguistic structure, a chronicle can be written in Latin or in vernacular, in prose or in verse. Here also belong the grammatical structures. On this occasion we need not speak about these structures which are not immediately visible in the text itself, but which can be marked by an editor, as e.g. the singular formulae in a charter. There might also be marks concerning quotations or other inserted texts in the given source, or interlinear textual additions, etc. The standardization and exchange of such editorial remarks correspond to a second level of text-treatment, not to standardization and exchange of the historical sources themselves. In addition there are still big problems of terminology. For the purpose of structural research orthographically normalized transcriptions are enough. Such transcriptions can also contain unmarked solutions of abbreviations and even emendations. This is the well-known model of the historic-diplomatic edition which will also be sufficient for a wide range of historical-diplomatic studies and for the history of literature, but only partly for linguistic interests. Concerning these historical studies there will not be any loss of information.

As to the phonemes linguists are interested in being able to distinguish all phonemes as far as possible, and therefore they do not agree with any orthographic normalization or other treatment of the text. The results of these principles are the also well-known philological editions which stick to the use of majuscles and minuscles and, if considered necessary, to the interpunction of the model in an overprecise way. As to the linguistic purposes there will be no loss of information, correct transcription provided.

As to the graphemes the paleographers are almost the only ones who also are interested in distinguishing even between graphemes which obviously do not signify different phonemes, e.g. if the dots on the ***i's*** are put or not, or if the Latin word *et* is fully spelt by the letters *e* and f, or indicated with commercial &, or represented by the Tironian sign which is similar to the figure 7. Paleographers are also interested e.g. in different kinds of characters in a given text, e.g. in the appearance of certain uncial letters in an inscription written in capital letters, in the use of certain ligatures, etc.

The paleographical transcription which partly meets such demands, is obviously the most rigid form of copying a given text. Only a few editions of inscriptions, however, will meet such demands. But we must admit that normally in a paleographical transcription there is not even a possibility to distinguish between medieval and modern writers immediately. Even the most precise paleographic transcription means an enormous loss of information in respect to paleographical criteria. We can only hope that in the future a computer-aided classification of scanned graphemes will be able to reduce this loss of information. Graph 2 illustrates the usability of the different kinds of text-representation for the various historical disciplines in a schematic way.



Often the editions of historical texts which really exist cannot be assigned easily to one or the other of the four types, but there are mixed forms. There are hybrid forms between a summary and a copied text. Even historical editions do not contain normalized forms only: the proper names

are usually given in the orthography of the model. Many editors of historical editions will correct obvious mistakes of the model only if the model is a copy and not an original. On the other hand even the historical editions contain some indications of mere graphic peculiarities, e.g. the *litiera elongata.* This is, by the way, also an example for the fact that certain graphic peculiarities are closely connected with the textual structure in general. Even the material on which the text is written (e.g. paper or marble) is not only connected with the graphemes, but also with the purpose and therefore with the content of the text. Finally we can remark that not only philological editions, but also several historical editions can distinguish between different graphemes for only one phoneme respectively, e.g. the two letters for *s* and the two letters for *z*.

Nevertheless on the whole we can say that summaries and normalized editions suffice for scientists who are interested in material contents, and that philological editions suffice for linguists. The paleographical transcriptions, e.g. carefully-made epigraphic editions, however, can only partly meet the demands of paleographers. Obviously the philological and paleographical transcriptions can also satisfy the demands of historians in the narrow sense of the word, but a linguist normally needs a philological edition at least, and finally the paleographer has a chance of finding some helpful information only in a paleographical transcription.

3. The Desirable Level of Text-Treatment

Now we want to find out which level is recommendable as a standard-level for the text-treatment. As we have seen, the level which could meet the widest-ranging demands is the paleographical transcription. But in this respect we are far away from any standard, and perhaps such a standard will never exist: The material is excessively diverging, the terminology far from uniform, and the possibilities of computer-aided examination of handwritings will increase and will continuously create new possibilities and new problems. On the other hand, most of the historians (in the narrow sense of the word) and most philologists will not want to do an enormous amount of extra-work for some paleographers. In any way, a common standard of text-input has to be more general, and the paleographers have to find out their own more specialized standards, e.g. for Roman epigraphs, for the inscriptions on medieval coins and seals and so on.

As to the linguists, literal transcription has already been proposed as a standard several times, which means more or less the level of the philological edition (6). But we run the risk that even such a solution would not work in a satisfactory way. Only recently great enterprises have been started which follow the historical-diplomatical tradition of normalized edi-

tion, and these efforts have been promoted by scientists who are very familiar with the discussion about standardization and also with the suggestions for literal transcription as a desirable standard. Generally historians do not do extra-work for philologists. They would have to transcribe the model first according to philological rules and after that to correct the text gained according to their own rules. There are also all the indispensable editions which already exist and which we can make machine-readable only afterwards. They will remain, of course, in their normalized forms, and it will often happen that a substitution of these old editions will not be possible. For that reason we have to realize that we might declare the literal transcription to be a standard as often as we like, but that our declarations would be useless. For after many years even the linguists will still have to work on the basis of normalized editions.

But not even the third level, that is the normalized historical edition, is suitable for a standard either. The rules for normalization and treatment of texts not only vary from language to language and from period to period, but also from country to country and even more from edition to edition according to its main purpose. It is certainly hopeless to decree a certain rule as a standard, after we have realized that the different results of the various rules of normalization cannot (or can only partly) be converted among themselves.

What is to be done? Neither the paleographical nor any form of the historical-diplomatical transcription can serve as a standard for the whole scientific community. Therefore we have to define the literal transcription as a desirable standard, but we also have to realize that beyond this standard the paleographers have to find their own more precisely defined standards, and we have to realize as well that the greatest part of the historians (in a wide sense of the word) do not come up to this standard. Under those conditions »standardization« can no longer mean uniforming the text-treatment itself, but ensure the comparability of the various transcriptions we have to cope with. This can be done by standardization of documentation.

4. The Documentation of Text-Treatment

As to the documentation of machine-readable texts, it is obvious that besides all technical details such a documentation has first of all to meet the same demands as the introduction of a conventional edition as follows: firstly indications concerning the source itself (author, text arranger or translator, title, genesis of the text respectively of its versions, language, linguistic peculiarities of the text respectively of its versions, special indications, e.g. of authenticity, purpose and importance of the source), se-

condly indications concerning the tradition of the source (location, e.g. of a manuscript, with codicological description, editions, special indications, e.g. of reactions in later texts), thirdly bibliography and finally indications concerning the edition itself (name of the editor respectively of the reviser, etc.). A documentation of machine-readable sources must, of course, also contain a number of technical details, first of all information on conventions concerning the characters and sequences used. As to continuous texts we also need a list of the characters which can appear within a word (or within a number), further more a list of the characters which separate the singular words and a list of the characters which separate the singular phrases from one another, occasionally a description of the editorial signs indicating additional remarks of the editor. It is also recommendable to give information on the user's rights and on the opportunities of obtaining further information.

Most of this information in the documentation can hardly be formalized, but a considerable part of it can and should be formalized. As to the treatment of text before the input the task of the documentation should not only be informing the user about the usability of singular details in the machine-readable text, but this part of the documentation should be able to put two things in effect: Firstly a linguist who tries to find certain orthographic particularities in various data-bases worked out according to different principles of normalization must be sure that the system can ignore all normalized forms or, possibly, can take in account only certain defined normalized forms (7). Secondly the system should be able to set up a list of the singular data-bases (or documents) taken in account, and the singular data-bases (or documents) ignored respectively excluded in connection with the question on hand. In this way the researcher, the linguist e.g., would be able to realize and to evaluate right away the wide or narrow basis of his results.

These opportunities will be of advantage especially if the given problem requests a checking of a great number of shorter documents which have been transcribed by different researchers or editors. We also know very well that certain data-bases contain »mixed« material treated according to different principles, e.g. original documents and copies.

In contrast to the introductions to conventional editions such a documentation must be formalized. Besides the possibility of generic rules such a documentation of text-treatment will consist of a sort of list. Over the listing of original forms and normalized forms additional examples or explanations in the documentation will be useful as well. The following simple examples refer to certain problems of the medieval Latin language and Latin characters.

Here belong: normalization of Latin endings according to the general medieval use (e.g. *abbatisse* and *magistre* instead of *abbaiissae* and *magistrae):*

»ae_« → »e_«
»ae.« → »e.«
»ae,« → »e,«

ignoring of accents (e.g. *a* instead of $\acute{a}$ and $\hat{a}$):

»á« → »a«
»à« → »a«
»â« → »a«

and general use of minuscules (e.g. *a m e n* and ... *communimus, statuentes* ... instead of *Amen* and ... *communimus, Statuentes ...):*

»A« → »a«
»B« → »b«
»C« → »c«

Nevertheless the first letter of each sentence must be a maiuscule (e.g. ..., *amen. Bertholdus episcopus* ... instead of ..., *amen, bertholdus episcopus* ...):

»._a« → «._A«
»._b« → «._B«

»?_a« → «?_A«
»?_b« → «?_B«

Furthermore the very first letter of certain terms must be a maiuscule (e.g. *Deo* instead of *deo):*

»_deus« → «_Deus«
»_dei« → «_Dei«
»_deo« → «_Deo«

There can also be normalization of *Uu, Uv, Vu* and *Vv* according to the phonetic significance (e.g. *Waherus* instead of *Uualterus* or *Vualterus,* but *Vulfingus* instead of *Uvlfingus):*

»_Uua« → «_Wa«
»_Uva« → «_Wa«
»_Vua« → «_Wa«
»_Vva« → «_Wa«

»_Uul« → «_Vul«
»_Uvl« → «_Vul«
»_Vvl« → «_Vul«

As soon as that is done a normalized edition can possibly exchange each remaining *u* and v according to the phonetic significance (e.g. *videlicet* instead of *uidelicet*, but *rubeus* instead of *rvbeus):*

»_ua«	→	«_va«
»_ue«	→	«_ve«
»_ui«	→	«_vi«
»_vb«	→	«_ub«
»_vc«	→	«_uc«

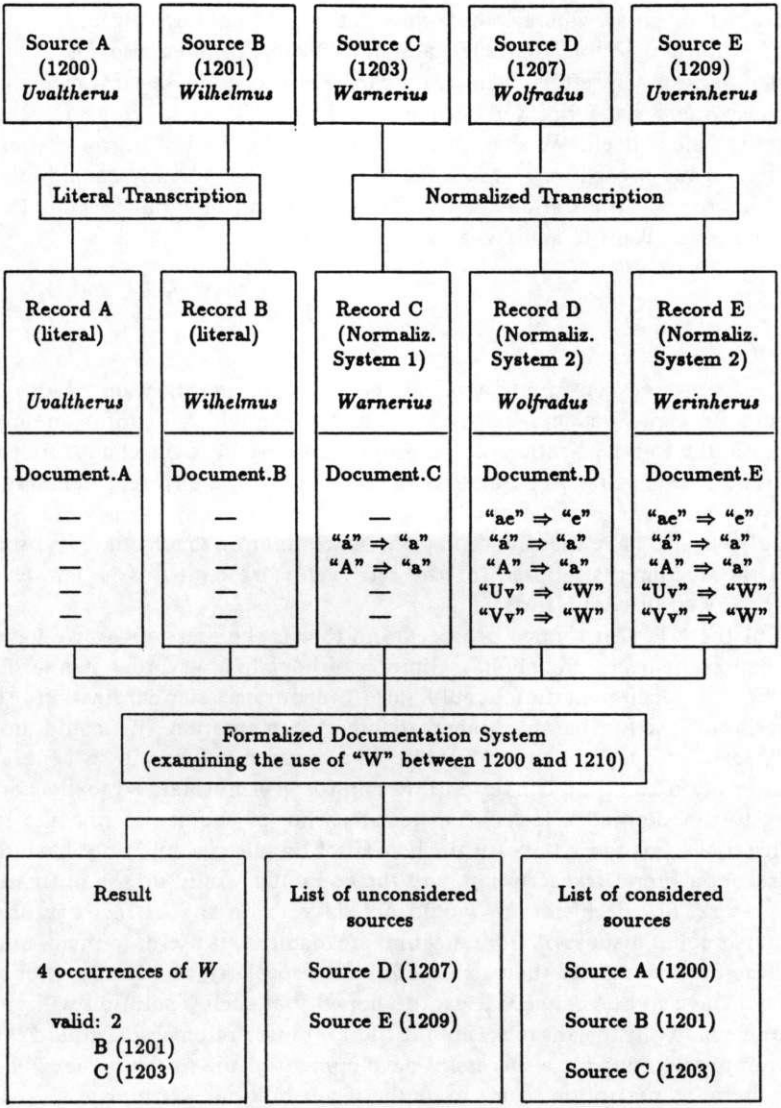
As to the massed consonants in the German texts of the fifteenth and sixteenth century the list will demonstrate, e.g. the reduction from *krojjt* to *krajt* and from *markght* to *markt:*

»fft«	→	»ft«
»kght«	→	»kt«

Some problems will remain, of course. The system has to learn to distinguish between proper names and other words, between words and Roman numerals, etc. The system also has to learn to find out, if a point in the text is a full-stop or an abbreviation symbol, etc. It is also obvious that some more information must be inserted in a suitable way in order to make such a system function perfectly, e.g. the documentation must also contain not only the date of origin of copies as far as this is possible, but also the date, when a forgery was written. Otherwise a researcher could easily be tricked, because the distinction between the pretended and the real date of origin could not be made.

But apart from these problems such a system could easily be able to ignore (on the basis of the given example) all letters *W* in corresponding normalized documents, even if a scholar wants to study the use of *Uu, V v, Vu, Vv* and *W* in a certain number of data-bases or documents. Graph **3** illustrates how in this case the documentation could (and should) work.

It must be admitted that these suggestions make the impression of causing much extra-work for researchers and editors. But we can considerably facilitate this extra-work. We only need such a list for every language or period. These primary list should contain every orthographic normalization which makes sense. On this basis we could easily create more special lists containing the rules of the large series of editions. We only have to leave out some not applicated rules. Normally also singular editors or researchers do not create new rules of transcription, but they simply adapt some existing rules. Therefore some easy adaptations will be enough for obtaining every particular list required.


Source A (1200) *Uvaltherus*
Source B (1201) *Wilhelmus*
Source C (1203) *Warnerius*
Source D (1207) *Wolfradus*
Source E (1209) *Uverinherus*
Literal Transcription
Normalized Transcription
Record A (literal) *Uvaltherus* Document.A
Record B (literal) *Wilhelmus* Document.B
Record C (Normaliz. System 1) *Warnerius* Document.C
“á” ⇒ “a”
“A” ⇒ “a”
Record D (Normaliz. System 2) *Wolfradus* Document.D
“ae” ⇒ “e”
“á” ⇒ “a”
“A” ⇒ “a”
“Uv” ⇒ “W”
“Vv” ⇒ “W”
Record E (Normaliz. System 2) *Werinherus* Document.E
“ae” ⇒ “e”
“á” ⇒ “a”
“A” ⇒ “a”
“Uv” ⇒ “W”
“Vv” ⇒ “W”
Formalized Documentation System (examining the use of “W” between 1200 and 1210)
Result
4 occurrences of *W*
valid: 2
B (1201)
C (1203)
List of unconsidered sources
Source D (1207)
Source E (1209)
List of considered sources
Source A (1200)
Source B (1201)
Source C (1203)

As to the old editions which only have been made machine-readable afterwards it can be recommended to use the first (and most complete) list as long as the principles of the editions have not been examined. In this way the researcher is best protected from being presented some data as presumed originals whereas these very data might be normalized.

It is most probable that such a standardized documentation is much easier to be put in effect than a standardized text-input itself, especially because there have not yet been established differing rules as for the text-treatment itself. We shall have to apply in the field of normalization an analogous procedure of the kind Mr. Thaller has already roughly developed for the translations of coding-systems and of input formats (8). We, however, want to achieve a different result.

5. Conclusion

And now we can summarize: an effort in standardization of documentation can - at least partly - close the gap which is unfortunately given in the standardization of text-input itself. We have the chance firstly to use all texts as far as possible corresponding to the different demands, and we can manage without doing too much embarassing extra-work. Secondly we also have the chance to get a list containing the documents used and the documents ignored for the retrieval-work, and this is done without any additional efforts.

But there is also a third prospect, and that is the reason why we have spoken at the beginning about a glimmer of hope in a way, that standardization of documentation could also encourage standardization of text-input itself. The machine-readable documentation-file could not only serve as documentation, but also as a command-file, and so a literal transcription could be converted into any form of normalized text according to pre-defined rules. The historians who, of course, do not like to transcribe a text twice only for the benefit of linguists, would only have to work out a literal transcription, and the computer would do the normalizing work. So the historians would not have to do any extra-work, the linguists could dispose of more machine-readable texts useful to them, and the sacrosanct rules of the various editions (especially historical editions) do not have to be changed. It can be hoped that such a solution will encourage many historians to accept the literal transcription as a standard for text-input. Hoping for a successful development of the software described the literal transcription seems to be the ideal solution even today.

So in the end, we have to ask ourselves: Who shall do this work, that is who shall develope the »converting-system«? It is not because I am a historian, that I answer: Let the philologists, and especially the linguists do

this work. I suggest this because the linguists should be those who are most interested in it. In any way I am convinced that a standardization will not function if the extra-work has not to be done by these researchers who benefit from the efforts. Moreover it is the linguists who command of the necessary knowledge in the first place. Certainly they will be assisted by interested historians, and I will also join them.

Notes

1 This article is an extended version of my contribution to the Cologne Computer Conference held on September 9th, 1988: *The Demands of the Historical Disciplines on Machine Readable Sources - and the Consequences for their Standardization.* The article deals with some very general problems, and I have had to decide in favour of an enormous bibliography or in favour of some essential bibliographical remarks. The latter solution seemed preferable.

2 See *Datennetze für die Historischen Wissenschaften? Probleme und Möglichkeiten bei Standardisierung und Transfer maschinenlesbarer Daten / Data Networks for the Historical Disciplines? Problems and Feasibilities in Standardisation and Exchange of Machine Readable Data,* ed. by F. Hausmann et al., Graz 1987.

3 See *Standardisation et Echange des Bases de Données historiques. Actes de la troisième Table Ronde internationale tenue au L.I.S.H. (C.N.R.S.) Paris, 15-16 mai 1987*, ed. by J.-Ph. Genet, Paris 1988.

4 That is also true of the wide-scale *Text Encoding Initiative* as it was presented by N. M. Ide and C. M. Sperberg-McQueen at the Cologne Computer Conference. See *Cologne Computer Conference. Cologne, September 7th-10th, 1988, Volume of Abstracts,* pp. E.6/3-E.6/4.

5 I must admit that my special field are charters and similar documents. But these particular texts contain both elements of continuous (narrative) texts and elements of quantitative historical material (administrative records), and therefore the charters must be considered as a suitable example for historical texts in general terms.

6 See M. Thaller, *Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften*, in: M. Thaller (ed.), *Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung* (= Historisch-sozialwissenschaftliche Forschungen 20), St. Katharinen 1986, pp. 9-30, here pp. 16-17; and especially L. Fossier and M. Parisse, *Aufnahme und Verarbeitung mittelalterlicher diplomatischer Texte,* pp. 78-81, here p. 81; M. Parisse, *Standardisierung und Austausch fortlaufender Texte,* pp.

194-196, here p. 195; Ch. Zinko, *Aufbereitung hethUischer Keilschrifttexte für eine computerunterstützte Verarbeitung,* pp. 248-257, here p. 255 (all in: Hausmann et al. 1987).

7 Data-base i.e. a collection of data.

8 See M. Thaller, *A Draft Proposal for a Standard Format Exchange Program,* in: Genet 1988, pp. 329-375.